Technical Disclosure Commons

Defensive Publications Series

April 24, 2017

Active Protection Of Voice-Activated Assistant Against Accidental And Malicious Activations

Sandro Feuz

Thomas Deselaers

Victor Carbune

ACTIVE PROTECTION OF VOICE-ACTIVATED ASSISTANT AGAINST ACCIDENTAL AND MALICIOUS ACTIVATIONS

ABSTRACT

System and method are disclosed to protect a voice-activated assistant against accidental or malicious activation. The system includes a device interacting with the cloud having an assistive device, which in turn is protected by another, protective assistant. The protective assistant detects incoming malicious (or spurious) actions triggered by a third party which tries to connect to the device through the assistant. The protective assistant then emits a signal that would perturb the third party signal, or prepares and issues “cancellation” commands to automatically cancel the action. The system may also notify the user via SMS / phone call / email / screen notification to further manually check the legitimacy of the detected malicious action. User experience is improved as accidental or unintended orders will not be executed by their assistants.

BACKGROUND

Currently, multiple companies have developed voice activated assistants. Some users may have multiple assistant devices from more than one company. These devices may be of different form factor, have different amounts of compute power or different other hardware features (such as good speakers or microphones).

Sometimes, an assistant is activated accidentally (this is a false positive of the hotwording mechanism). However, assistants could also possibly be activated maliciously - e.g. an "attacker" trying to make an assistant of one user do something bad. Apps may be built with the intent of purposely emitting malicious intents through an assistant to obtain an illegitimate advantage. This may include botnets, sending email spam, search engine spam, or click-spam.

DESCRIPTION

A system and method are disclosed that aim to cancel or prevent issued actions that could potentially lead to harmful or unintended decision by an assistant when another assistant or a known source sends a request. The system includes a user interacting with the cloud through an assistant device, which in turn is protected by another assistant as shown in FIG. 1.

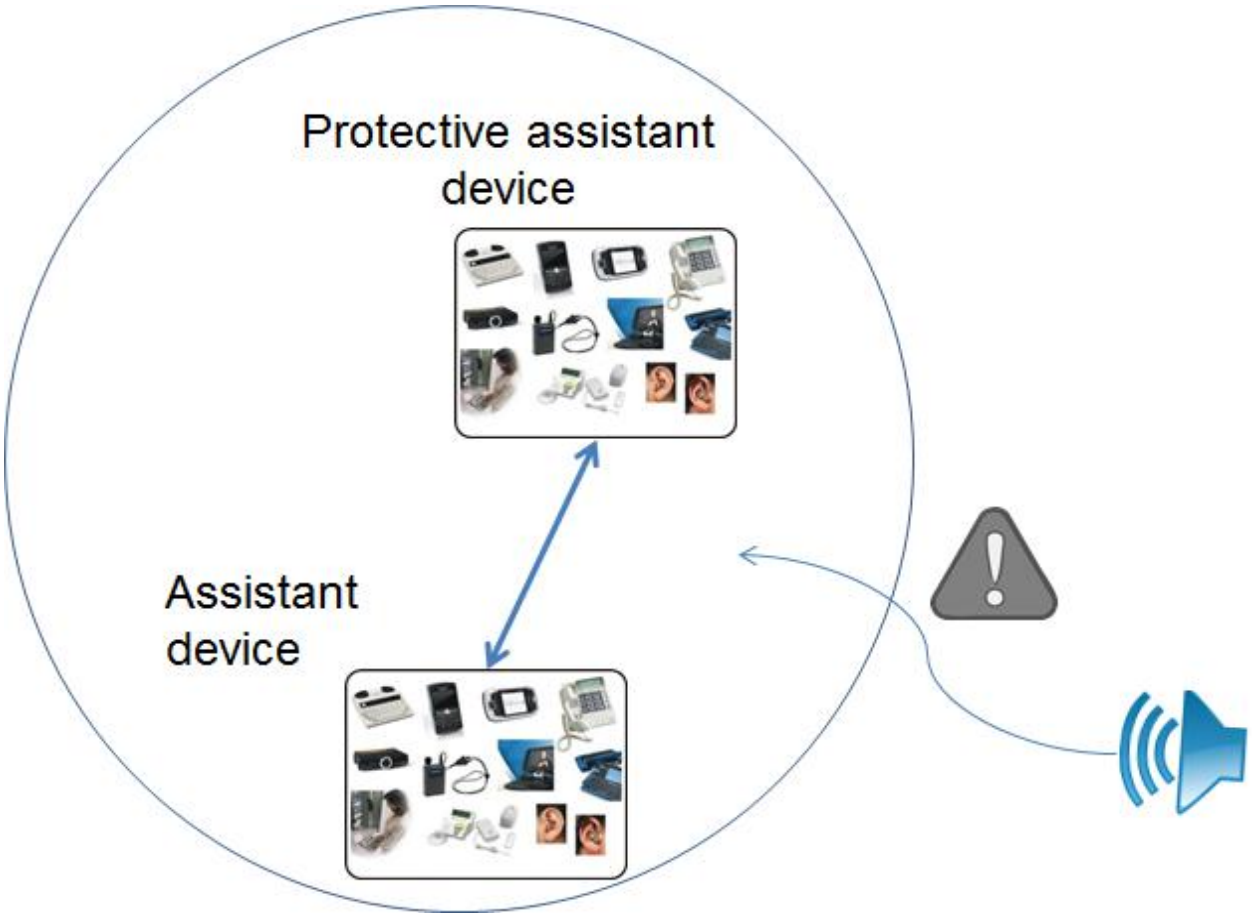

FIG. 1: Protective assistant against malicious voice-activated attacks

The method of protecting an assistant from incoming malicious content using the system of FIG. 1 is shown in FIG. 2. A third party or some other source tries to connect to the system through the assistive device. The protective assistant receives the connection attempt in parallel, and detects any malicious (or spurious) action emitted by the third party source. If malicious

content is detected, the protective assistant emits a signal that would perturb the third party signal, or prepares and issues "cancellation" commands to automatically cancel the action.

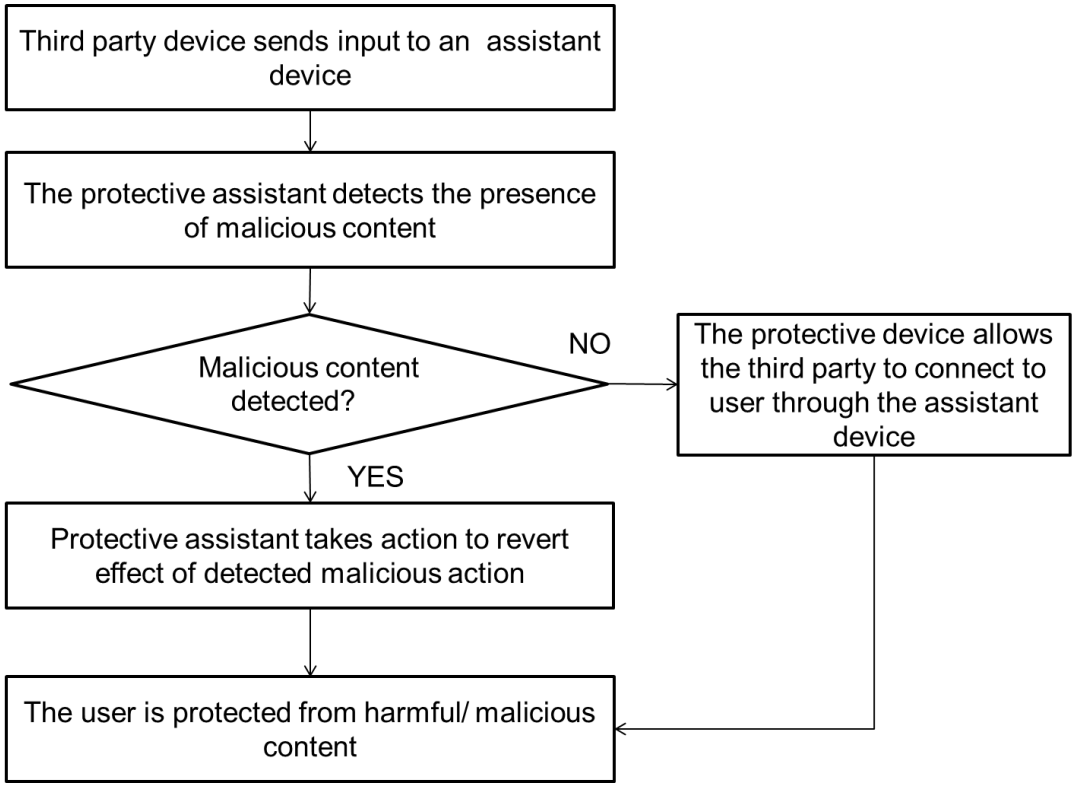

FIG. 2: Method by which a protective device protects an assistant device

Various assistants may have different capabilities and in some cases, an assistant device may be in a good situation to prevent another assistant from making a mistake, which the other may fail to detect and prevent itself. For example, if an assistant has network access and the other has no network access, or if one is on a very low powered device (a cell phone or a watch) and the other is a powerful home-assistant device.

The user of the protective assistant could configure what it considers a malicious or invalid query. Depending on the hardware capabilities of the protective assistant, those settings of what is considered a malicious query could vary. For instance, if the protective assistant has the capabilities to identify speakers by their voice, it could be configured so that only specific

speakers can issue commands for the other assistant. Also, malicious or rather unintended actions against an assistant might come from recorded queries such as a TV or video or the like. The protective assistant may have a database of fingerprints of pre-recorded occurrences of hotwords on videos or TV shows or the like. Suppose, for the purpose of example, the protective assistant is assistant A, and it has speaker identification, and further suppose there is an assistant B which does not have speaker identification. The protective assistant A could then be configured so that only a specific set of speakers are allowed to issue commands to assistant B (even though assistant B cannot distinguish them by itself). Whenever assistant A hears the hotword for assistant B, it performs the speaker analysis and checks whether the speaker is in the set of people allowed to issue commands against assistant B. If not it takes actions to counteract the query.

Taking actions to perturb the the malicious action that may be detected involves the following steps. The protective assistant tries to automatically disturb the action/signal from the third party by emitting a signal in real time that perturbs or cancels the incoming signal. If the same device emits the action, the protective assistant simply turns the output stream to not be understandable anymore. If another device is emitting such an action, then the protective assistant may emit an additional command to perturb it. Alternatively, the system may immediately notify the user via SMS / phone call / email / screen notification to further manually check the legitimacy of the detected malicious action.

Alternatively, the user is protected against the maliciously emitted command by further processing and understanding the effects of the command to a finer grained level. For example, the protective assistant may have recognized that the listening assistant is a third-party assistant that has account details and ordering rights on a particular product (e.g. online shopping app

store). The protective assistant prompts the user to quickly open the app and cancel the order immediately after the malicious command was processed by the assistant. Depending on the permissions on the protective assistant, the user's credit card could be locked temporarily to protect against the unauthorized purchase.

A protective assistant that responds well to a situation can prevent another assistant from making a mistake during such situations, which the assistant device fails to detect. The disclosed method would lead to improved user experience as accidental or unintended orders will not be executed by their assistants.